Final Business Report: Car Insurance Claim

**Introduction:**

**a) Defining Problem Statement:** The objective of this project is to build a “Risk Analytics model” to understand the renewal potential and claim propensity of Existing Customers under Personal Auto Insurance Lines.

**b) Need of the Project:** The Insurance industry is immensely data intensive. Historically, their data has been largely fragmented and underutilized. Also, insurance industry goes through a lot of combining structured and unstructured data, which enables the industry to generate powerful insights. With an incredible amount of data flowing in from multiple new digital channels, the insurance industry is undergoing a paradigm shift in the way they function – right from product planning to pricing, introduction, marketing, customer self-service and claim processing.

**c) Understanding Business/Social Responsibility:** As of 2016, there were over 268 million [registered vehicles](https://www.statista.com/statistics/183505/number-of-vehicles-in-the-united-states-since-1990/) on the roads in the United States. In 2015, there were 32,166 [fatalities](https://www.statista.com/statistics/191499/vehicle-crashes-by-severity-in-the-united-states-since-1991/), 1,715,000 injuries and 4,548,000 car crashes which involved property damage. So, while many of us feel secure in our vehicles, the statistics indicate the importance of automobile insurance and in most cases, auto insurance is required by law. Auto insurance is important because it not only covers any physical damage that may occur in an accident, but also any damage or injury that might be caused because of a vehicular accident or which may be done upon oneself or one’s vehicle by another vehicle or accident – a falling tree for example (Car insurance in the U.S. - Statistics & Facts)

**Data Report:**

1. **Source of the Data:** The dataset is collected during the year 2015-2016 for an auto insurance company. Cars manufactured through the year 1957-2000 are recorded in this data set.
2. **Visual Inspection of the data:** The data consists of 127 columns and 14178 rows. This data also contains several missing and incorrect values.
3. **Understanding of attributes:** “ClaimStatus” is the dependent variable. This is the categorical variable. It Indicates whether the policy holder has made a claim or not. 1 indicates a claim and 0 indicates no claim. Rest of the variables is independent variables, which we will analyze to predict our prediction model.

This Project is a challenging one, as it deals with a huge unbalanced data. I will be exposed to various ML techniques approach and able to find out the appropriate insights and recommendation. I will be enriched with real life problem handling and coming to the solutions.

**EDA - Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. - Both visual and non-visual understanding of the data.**

We are having the dataset which

* consists of 127 variables and 14177 observations.
* The Dependant variable is ‘ClaimStatus’, which is a categorical variable, contains 0 & 1 as ‘0’ indicates ‘no claim’ & ‘1’ indicates a claim.
* The data shows out of 14177 records only 778 records have claims which is about 5%.
* As the data contains majority parameters, which are non-significant for the response variable, we omitted all those variables to get better result for predicting.
* Final working dataset has 14177 observations and 22 independent variables. Different variables and the individual relation are shown below.

ClaimStatus**:** It Indicates whether the policy holder has made a claim or not. 1 indicates a claim and 0 indicates no claim. Initially the data type of this variable is numerical. But as the values are “1” and “0”, we need to convert it as Factor or categorical variable.

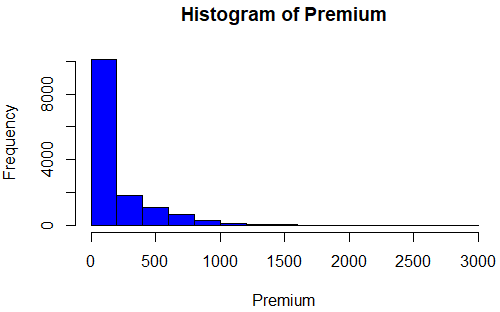
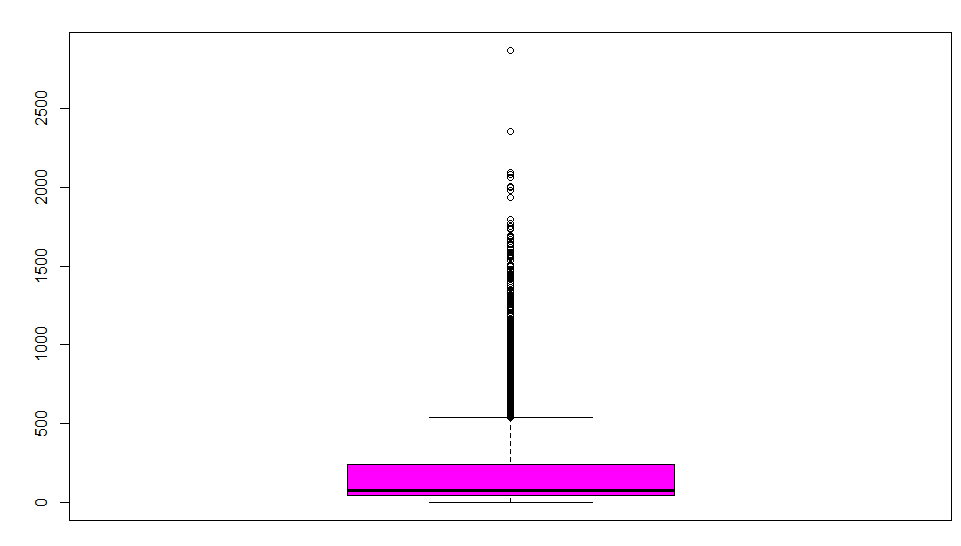
|  |
| --- |
| ClaimStatus |
| 0 1 |
| 13399 778 |

This means that 778 of total 14177 policy-holders have claimed the insurance and 13399 haven’t. There are no missing values or outlier in this field.

ClaimFrequency: Gives the number of claims claimed. The initial data type is numerical.

|  |
| --- |
| ClaimFrequency |
| 0 1 2 3 4 5 |
| 13399 580 144 36 16 2 |

Out of 778 total claims, 2 Claims were claimed 5 times, 16 claims for 4 times, 36 claims for 3 times, 144 claims for 2 times and finally 580 claims for single time. This claim frequency suggests maximum claims have been done for single or double time.

Premium: Premium in $1000. There is outlier. Maximum premium are bought under 50,000$.

Billing\_Term: How often the premium is paid, i.e. once a year = 1, Three times in a year = 3, or 6 times in a year = 6. This is also a categorical variable. But initially the data type is numerical. So we need to convert it to categorical variable. There are no missing values or outlier. The distribution is:

|  |
| --- |
| Billing\_Term |
| 1 3 6 |
| 6724 482 6971 |

Renewed: Indicates whether the policy has been renewed or not. 1= Renewed, 0 = not renewed. This is a categorical variable. But initially the data type is numerical. We need to convert it to categorical variable. There is no missing values or outlier. The distribution is such that policy renewal has almost equal ratio of zeros and ones.

|  |
| --- |
| Renewed |
| 0 1 |
| 7207 6970 |

DOB1 to DOB5: Date of Birth of the main policy holder/ main driver and followed by second driver till fifth driver. The data format is different from each other. All DOBs need to be derived to a single format.

$ DOB1: POSIXct[1:14177], format: "1947-10-26" "1958-07-29" "1961-06-04"

$ DOB2: POSIXct[1:14177], format: "1987-09-30" "1936-08-25" "1955-11-28"

$ DOB3: POSIXct[1:14177], format: NA "1965-10-18" NA ...

$ DOB4: chr [1:14177] NA NA NA NA ...

$ DOB5: chr [1:14177] NA NA NA NA ...

Number\_of\_Driver: Count of the number of drivers in the policy. The initial data type is numerical. We need to convert to the categorical variable. The number of drivers ranges from 1 to 5. The distribution of this variable is below. Majority of the policy are registered with 1 and 2 Drivers.

|  |
| --- |
| Number\_of\_Driver |
| 1 2 3 4 5 |
| 8624 4636 733 150 34 |

AgeUSdriving\_1 to AgeUSdriving\_5: How long the driver has been driving from the Driver 1 till Driver 5. This sectors are all of numeric datatype.

$ AgeUSdriving\_1: num [1:14177] 62 48 45 32 43 48 46 59 29 32 ...

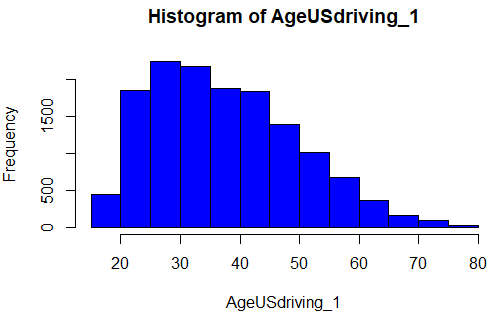
$ AgeUSdriving\_2: num [1:14177] 19 40 50 30 0 22 0 58 35 0 ...

$ AgeUSdriving\_3: num [1:14177] 0 41 0 0 0 0 0 22 0 0 ...

$ AgeUSdriving\_4: num [1:14177] 0 0 0 0 0 0 0 28 0 0 ...

$ AgeUSdriving\_5: num [1:14177] 0 0 0 0 0 0 0 30 0 0 ...

AgeUSdriving\_2 to AgeUSdriving\_5 have outliers. AgeUSdriving\_1 has majority of drivers aged between 25-35 years.



Amendment: Number of changes made to the policy during the year (may it is: No. of changes made to the policy till date from the date of buying). The initial Data type is numerical. We need to convert it to categorical. The distribution is as below.

|  |
| --- |
| Amendment |
| 0 1 2 3 4 5 6 7 8 10 |
| 13622 346 135 42 14 11 3 1 1 2 |

Majority of the claims hasn’t changed their policies in a year.

CoverageLiability: The three numbers represent (in the $ thousands) the liability limits for per-person bodily injury, bodily injury for all persons injured in any one accident, and property damage liability. Coverage liability represents the state's financial responsibility law i.e. the minimum requirement. the first figure is the amount to be paid for injuries per head, the second figure is the amount for all injuries and the third is the amount per vehicle damaged. For example, say you live in Ohio and hold the minimum amount of coverage, which is 25/50/25 (raised in December 2013). This means that the minimum liability limits in this state are $25,000 for injuries to one person, $50,000 for all injuries incurred and $25,000 for property damage for one vehicle in an accident.

|  |
| --- |
| CoverageLiability |
| 20/40/15 25/50/25 30/60/25 None |
| 7061 7077 38 1 |

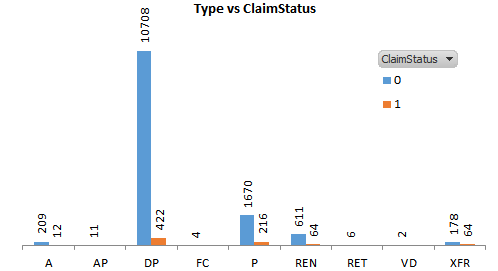
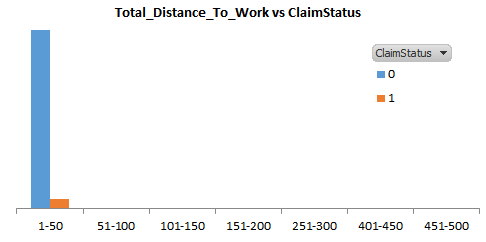
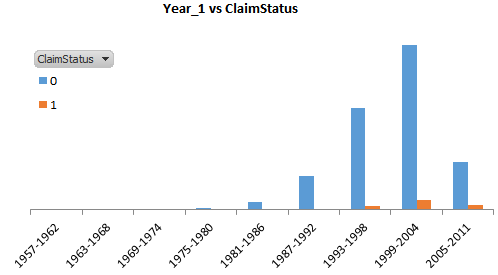
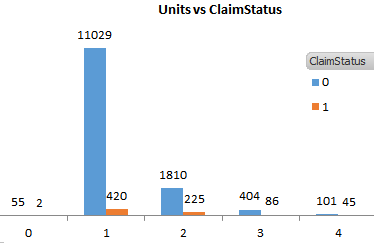
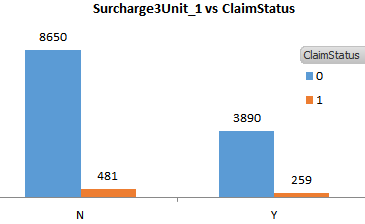
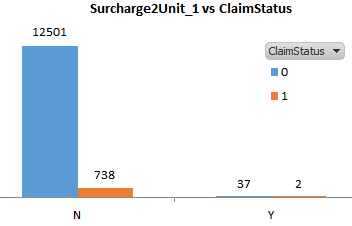
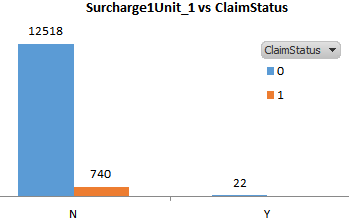
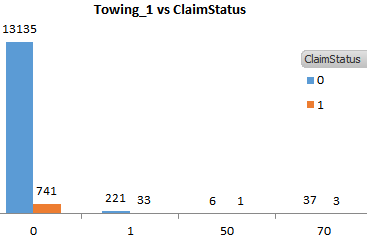
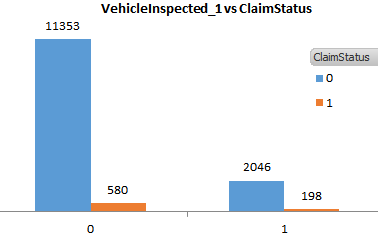
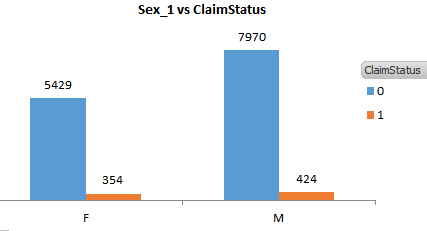
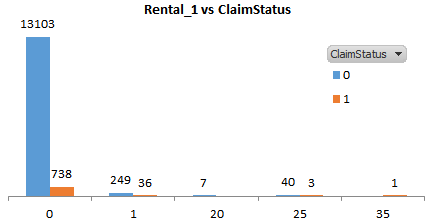
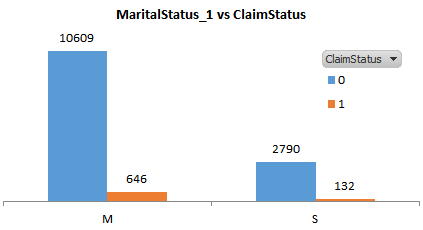
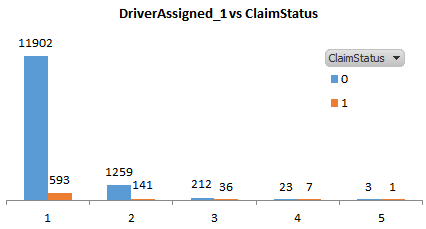
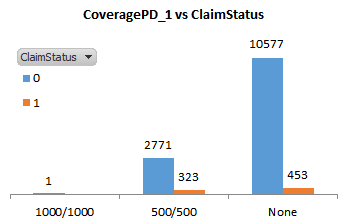
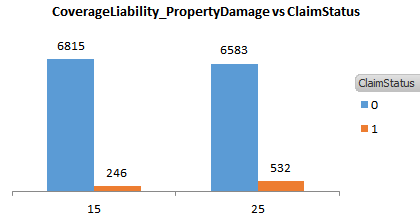
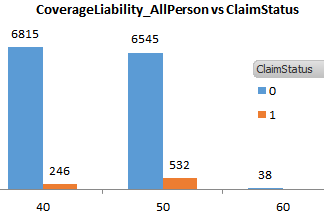
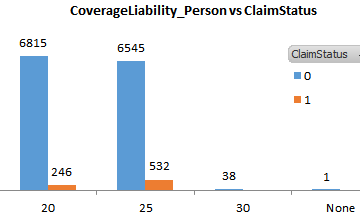
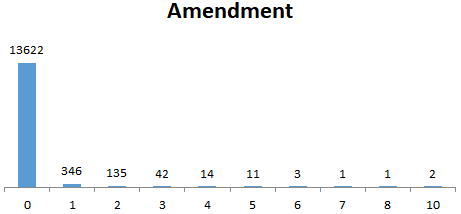
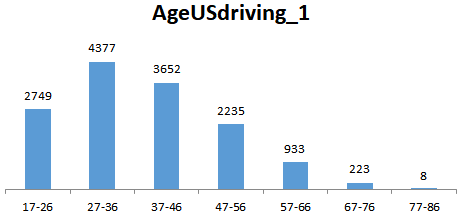
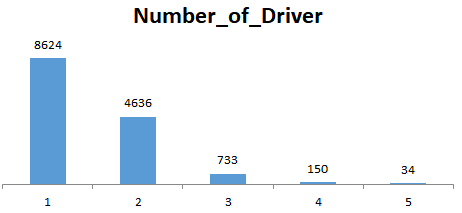
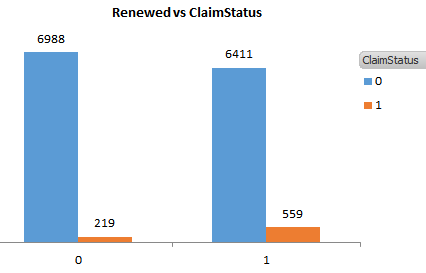
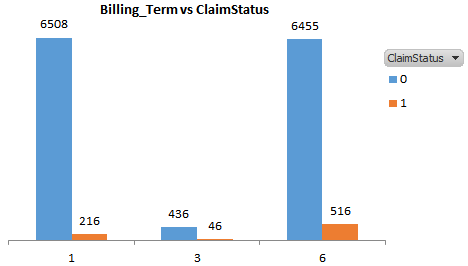
CoverageMP: MP stands for Medical payments coverage pays the reasonable expenses an insured person incurs for medical and funeral services within three years of an accident. As only 1 observation having data, we definitely need to drop this variable for further analysis.

|  |
| --- |
| CoverageMP |
| 535 None |
| 1 14119 |

CoveragePD\_1: PD is an abbreviation used in the car insurance industry. PD stands for Property Damage. It is a type of liability coverage and is part of the foundation of all state required auto insurance laws. Property damage insurance covers any damages to someone's property. This could mean a mailbox, someone's car, or even personal property in the person's car that was damaged as a result of the collision. PD is stated as two parts here, where the first part is the maximum payment that will be made for a single property in a single accident and the second number is the maximum that will be covered for all properties damaged in the accident. For example, the policy might show that you have Property Damage coverage of $25,000 per property, with a maximum of $50,000 per accident. Followed by CoverageUMPD & CoveragePIP\_CDW distribution are also shown.

|  |
| --- |
| CoveragePD\_1 |
| 1000/1000 500/500 None |
| 1 3094 11030 |

|  |
| --- |
| CoverageUMPD |
| Accepted None |
| 316 13855 |



Yearly Billing Count is higher, however, people who pays 6 times a year are likely to buy more claim

People who renew policy in a year, are likely to buy more claim

Single driver count is majority here

Age level between 27-46 years are the major driver age recorded

Usually majority of the customer do not change their policy in a year.

Coverage Liability has 3 segments, which we separated by definition. It can be seen that, the customer who have maximum limits for Per person injury, bodily injury for all persons and property damage liabilities as $25000,$50000 & $25000 respectively, are likely to claim more times.

453 customers who haven’t claimed for property damage bought claims. However, 323 people who recovered property damage amount have bought insurance claim

Mostly single driver assigned policy holders have more claims than other categories

It is observed that Claim amount is higher for the married persons

Majority cars are private. Also maximum claim comes from the private category cars only, not from rental category

Claim is irrespective of gender, both the male and females customers have similar ratio of claims

Majority of the vehicles not inspected, as in majority case there are very less damage to the car.

Similarly, very few vehicles needed towing due to some damage or problem.

Surcharges are very less common for the vehicles. However, third surcharge of the first vehicle is recorded.

Majority of the claim in this segment comes from where 1 or 2 number of vehicles are registered under policy.

Year of the manufacture of the first vehicle has not much impact on claim. However, where the manufacturing year is between 1999 to 2004, claim are relatively high.

Distance to work is not much impactful for claim. Claims are only there where distance is between 50kms.

DP & P type of the insurance have more claims comaparatively.

|  |
| --- |
| CoveragePIP\_CDW |
| 2535 2569 None |
| 1. 90 14036 |

|  |
| --- |
| CoverageUMBI |
| Accepted None |
| 316 13855 |

DistanceToWork\_1 to DistanceToWork\_5:

For Driver 1 & 2 distance to work for most of the driver is within 100 Miles and for 9 of the drivers it exceeds 100 miles. For Driver 3 most of the driver distance to work as in range 0-5 miles. Driver 4 & 5 covers miles in range 0-20 Miles and 0-5 Miles respectively.

DriverAssigned\_1: Count of drivers assigned to the first vehicle. 1 to max 5

|  |
| --- |
| DriverAssigned\_1 |
| 1 2 3 4 5 |
| 12495 1400 248 30 4 |

Engine\_1: Engine specification size in litres for the first vehicle. This field has lot of unclean data. We need to clean it.

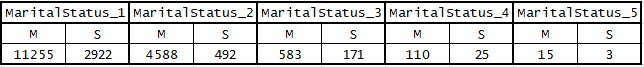
ExcludedDriverName\_01 to ExcludedDriverName\_20: First to twentieth person declared as an excluded driver respectively. As which customer is going to claim insurance has nothing to deal with the excluded drive names, thus these fields can be omitted in order to clean the total dataset.

GaragedZIP\_1 & Zip: There are two ZIP code fields in the data. The first variable indicates Zip code of the place where the first vehicle is parked. And the 2nd variable indicates Zip code of the insurance holder. These two fields also have nothing to contribute in predicting the claim status, these may be dropped.

We may also drop Relation\_1 field as there is no value except “Self”.

“CancellationType”: this field has only 320 values, thus this variable also needs to be dropped.

MaritalStatus\_1 to MaritalStatus\_5: Marital Status of the first to fifth driver. M - Married or S – Single

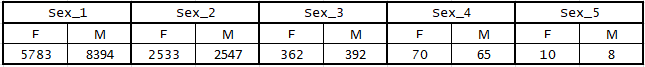


Occupation\_1 to Occupation\_5: Occupation of the first to fifth driver.

Rental\_1: first vehicle (If rental is allowed). Converting this variable from numerical to factor.

|  |
| --- |
| Rental\_1 |
| 0 1 20 25 35 |
| 13841 285 7 43 1 |

Sex\_1 to Sex\_5: Gender of the first to fifth driver M - Male, F – Female.



Surcharge1Unit\_1 TO Surcharge3Unit\_1: First to third surcharge for the first vehicle. Y - Yes, N- No

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Surcharge1Unit\_1 | | Surcharge1Unit\_2 | | Surcharge1Unit\_3 | |
| N | Y | N | Y | N | Y |
| 13258 | 22 | 13239 | 39 | 9131 | 4149 |

Towing\_1: first vehicle Towing and labor cost coverage is an optional coverage that you can add to your car insurance that typically protects you against some of the costs and hassles associated with common roadside breakdowns like dead batteries, flat tires or even an embarrassing lockout. (Some insurers may automatically fold this coverage into their policies, so be sure to ask.). Majority of the policy holder did not add this coverage.

|  |
| --- |
| Towing\_1 |
| 0 1 50 70 |
| 13876 254 7 40 |

Units: Number of vehicles covered in the policy. Majority of policy holder registered one vehicle under the insurance.

|  |
| --- |
| Units |
| 0 1 2 3 4 |
| 57 11449 2035 490 146 |

VehicleInspected\_1: first vehicle inspected. 1 - Vehicle was inspected, 0 - Vehicle was not inspected

|  |
| --- |
| VehicleInspected\_1 |
| 0 1 |
| 11933 2244 |

ViolPointsDriver: n-th time the m-th driver has scored violation point. From ViolPoints1Driver\_1 to ViolPoints8Driver\_5, majority of the observations are 0, thus these set of variables can be dropped.

Year\_1: Year of manufacture of the first vehicle. There are 22 records which are erroneous. The valid range of the manufacture year is from 1957 to 2011.

Total\_Distance\_To\_Work: Total Distance to work of all the drivers combined. This variable also can be dropped. As there is no relation of the distance covered with respect to claim.

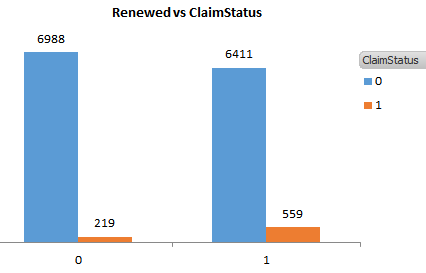
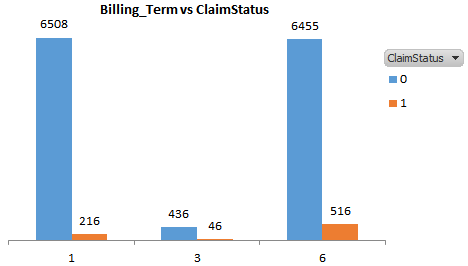
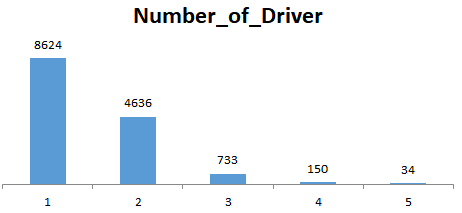
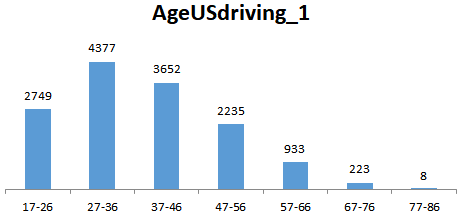
NoLossSigned: Whether statement of No loss has been signed or not. 1 - yes and 0 – No. This variable also can be dropped as signed observation is only 653.

|  |
| --- |
| NoLossSigned |
| 0 1 |
| 13524 653 |

Type: Different types of auto insurance viz, A, AP, DP, FC, P, REN, RET, VD, XFR. Distribution is as below.

|  |
| --- |
| Type |
| A AP DP FC P REN RET VD XFR |
| 221 11 11130 4 1886 675 6 2 242 |

Below are the plots shown for the relations of the different contributors to the response variable.

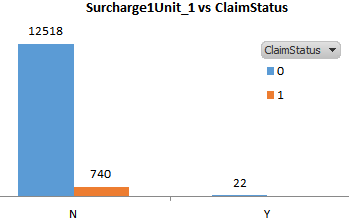
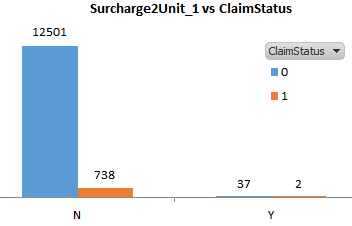
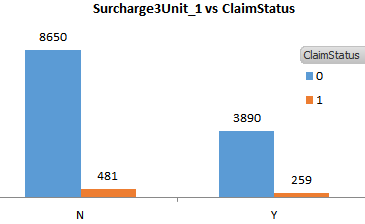


Age level between 27-46 years are the major driver age recorded

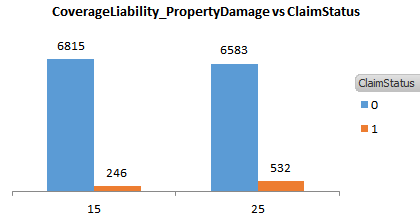
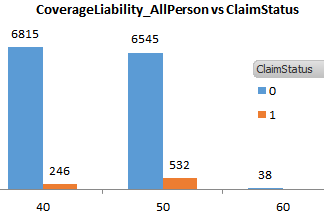
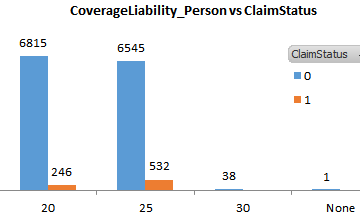
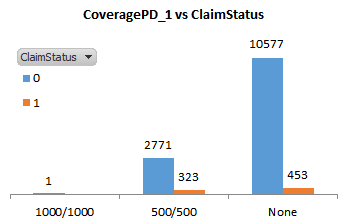
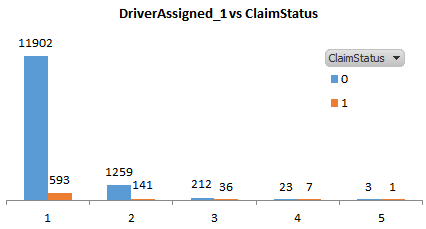
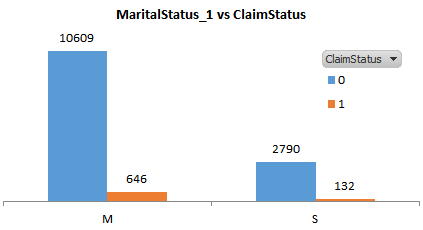
Yearly Billing Count is higher, however, people who pays 6 times a year are likely to buy more claim

People who renew policy in a year, are likely to buy more claim

Single driver count is majority here



Surcharges are very less common for the vehicles. However, third surcharge of the first vehicle is recorded.

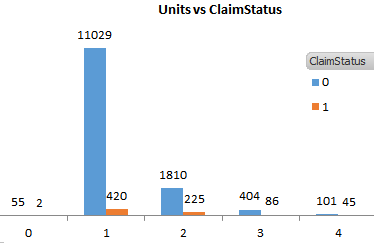
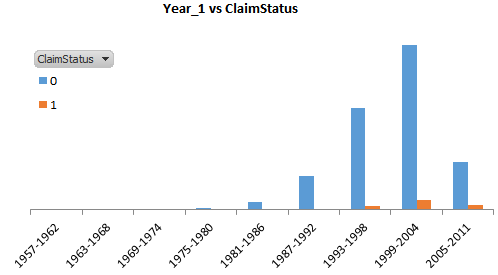
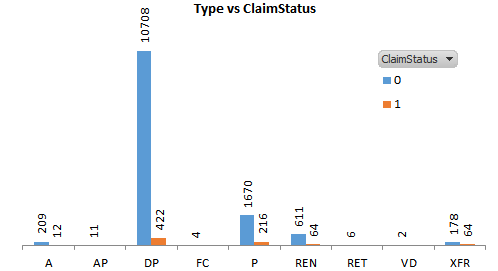
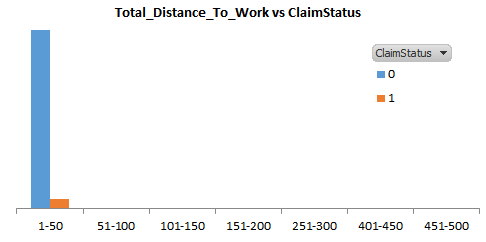


453 customers who haven’t claimed for property damage bought claims. However, 323 people who recovered property damage amount have bought insurance claim

Mostly single driver assigned policy holders have more claims than other categories

It is observed that Claim amount is higher for the married persons

Coverage Liability has 3 segments, which we separated by definition. It can be seen that, the customer who have maximum limits for Per person injury, bodily injury for all persons and property damage liabilities as $25000,$50000 & $25000 respectively, are likely to claim more times.

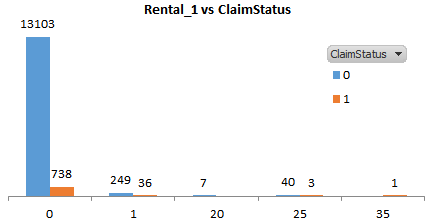
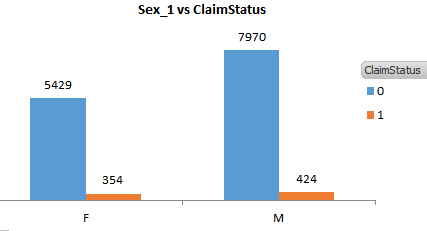
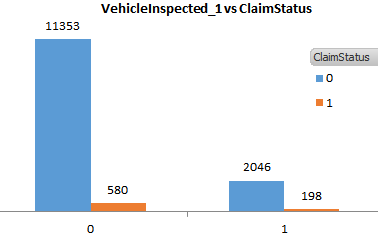
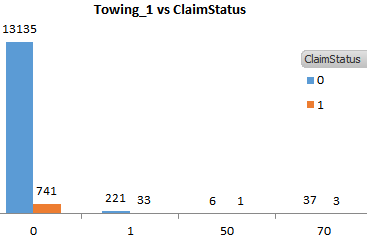


DP & P type of the insurance have more claims comparatively.

Distance to work is not much impactful for claim. Claims are only there where distance is between 50kms.

Year of the manufacture of the first vehicle has not much impact on claim. However, where the manufacturing year is between 1999 to 2004, claim are relatively high.

Majority of the claim in this segment comes from where 1 or 2 number of vehicles are registered under policy.



Similarly, very few vehicles needed towing due to some damage or problem.

Majority of the vehicles not inspected, as in majority case there are very less damage to the car.

Claim is irrespective of gender, both the male and females customers have similar ratio of claims

Majority cars are private. Also maximum claim comes from the private category cars only, not from rental category

**Data Cleaning and Pre-processing:**

There were missing values like

> colSums(is.na(mydata))

CoverageLiability\_AllPerson

1

CoverageLiability\_PropertyDamage CoveragePD\_1

1 52

CoveragePIP\_CDW CoverageUMBI

50 6

Surcharge1Unit\_1 Surcharge2Unit\_1

897 899

Surcharge3Unit\_1

897

For these missing values I have treated them with **Median and Mode imputation method** and successfully imputed all the missing values.

> colSums(is.na(mydata))

CoverageLiability\_AllPerson

0

CoverageLiability\_PropertyDamage CoveragePD\_1

0 0

CoveragePIP\_CDW CoverageUMBI

0 0

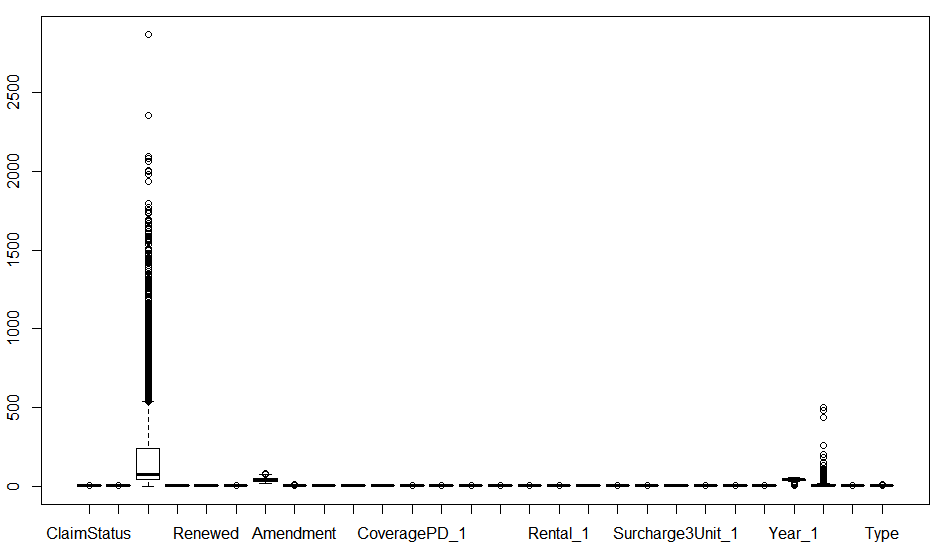
Surcharge1Unit\_1 Surcharge2Unit\_1

0 0

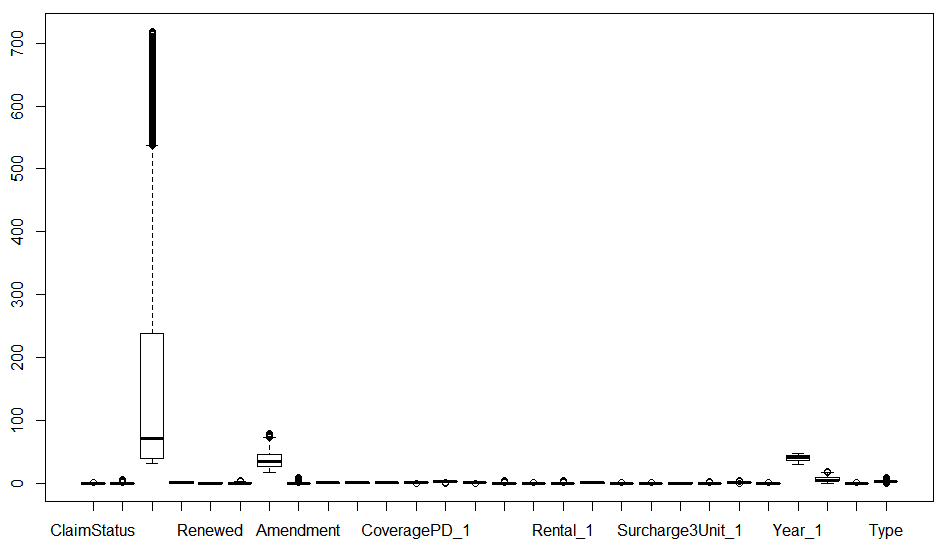
Surcharge3Unit\_1

0

Also there were outliers in a few variables.



So after treating the outliers with the help of squish() from library scales, I can see the overall outlier status as,



**Model building:**

I have implemented total 3 models: 1) Logistic Regression 2) Random Forest and 3) Naïve-Bayes.

The reasons for each selection are furnished below.

Naïve-Bayes Used due to:

1) Maximize the posterior probability given the training data.  
2) Naïve-Bayes classifier assumes that the effect of the value of a predictor variable on a given class is independent of the values of other predictors. This assumption is also called class conditional independence

Random Forest Used due to:

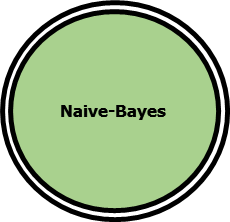
Logistic Regression models might be lacking in the accuracy, for which ensemble method such as Random Forest is adequate to boost the model’s accuracy

Logistic Regression Used due to:

1) To identify the significant variables which drive the Claim Status

2) Since the dependent variable is categorical variable, thus binary class of logistic regression required

3) Provides enhanced interpretability to identify various nuances of data



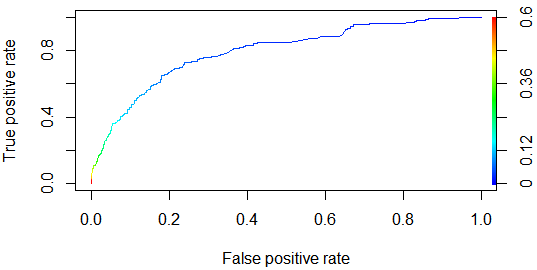
**Model validation:**

1. **Logistic Regression:** After implying final logistic model using the most significant variables, I got result as,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model3** | **Df** | **Deviance** | **Resid. Df** | **Resid. Dev** | **Pr(>Chi)** |  |
| Premium | 1 | 355.86 | 11339 | 4465 | < 2.2e-16 | \*\*\* |
| Billing\_Term | 2 | 32.93 | 11337 | 4432.1 | 7.08E-08 | \*\*\* |
| Renewed | 1 | 118.53 | 11336 | 4313.6 | < 2.2e-16 | \*\*\* |
| CoverageLiability\_person | 2 | 67.17 | 11334 | 4246.4 | 2.59E-15 | \*\*\* |
| Units | 4 | 115.74 | 11330 | 4130.6 | < 2.2e-16 | \*\*\* |
| Type | 8 | 22.14 | 11322 | 4108.5 | 0.004659 | \*\* |

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression\_Train | FALSE | TRUE | Accuracy |
| 0 | 10708 | 11 | 94.47 |
| 1 | 616 | 6 |  |

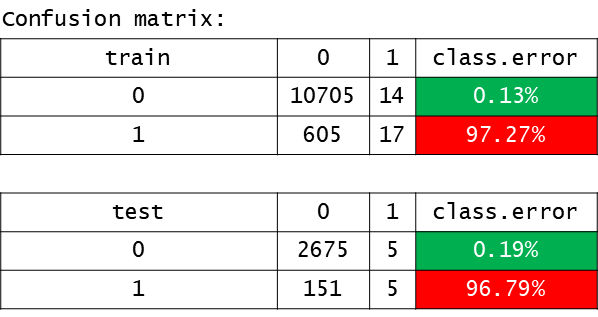
|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression\_Test | FALSE | TRUE | Accuracy |
| 0 | 2680 | 0 | 94.68 |
| 1 | 151 | 5 |  |

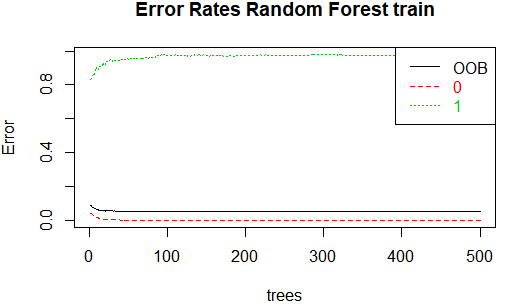
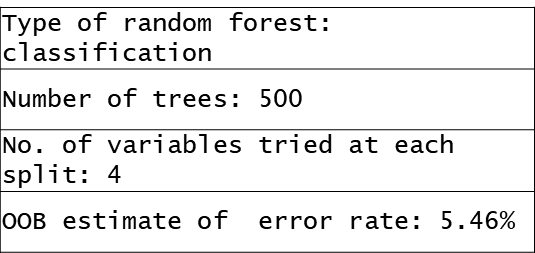


Though the accuracy comes to be ~94%, this model fails to predict the true-positive components. This also can be found from the AUC curve. So only Accuracy measurement is not enough to conclude.

Now to check any further betterment I will move forward to the Random Forest Model.

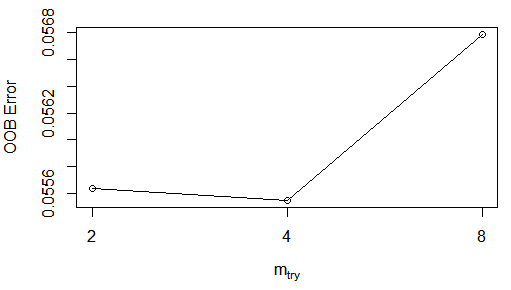
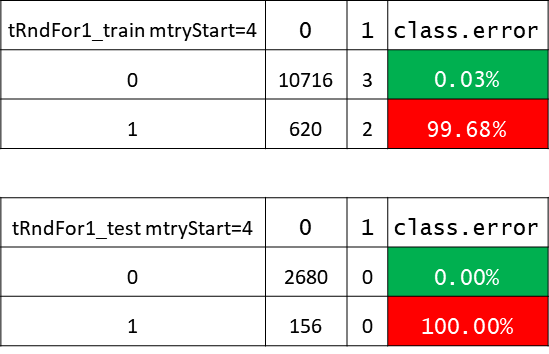
1. **Random Forest:** After implying normal Random Forest model using the most significant variables, I got result as,

****



Though the class I error comes to be as low as 0.19%, class II error increased hugely to approx. 97%. By this model I can predict only 17 persons who are going to claim out of 622 persons who actually claim (train data) and only 5 persons who are going to claim out of 156 persons who actually claimed (test data). Let’s see by tuning this model for further improvement.

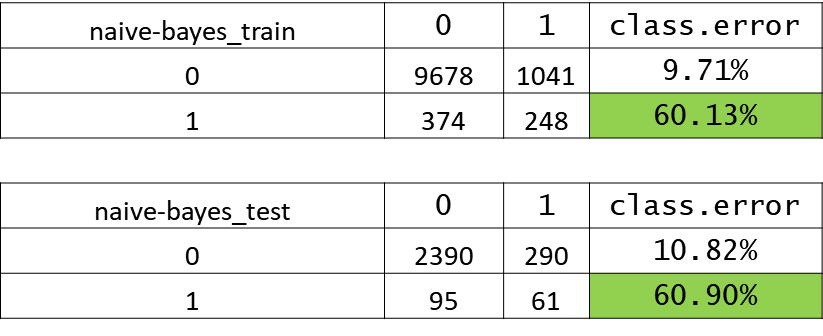
**Tuned Random Forest:** After implying final tuned Random Forest model using the tuned parameter, I got result as.



So, I can see by tuning mtry to 4, I got further improvement in predicting the persons who not going to claim (~100%), but again received highest error in predicting the persons who going to claim (Specificity~0%).

Now I will check with Naïve-Bayes Model for further overall performance.

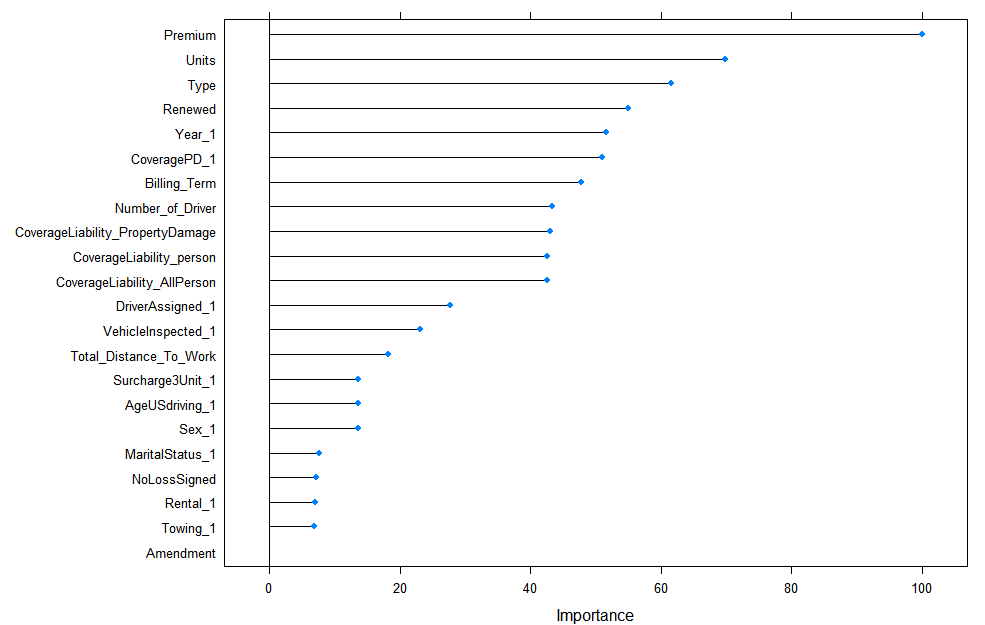
1. **Naïve Bayes**: After implying Naïve-Bayes model, we got result as:



By this model, I can find 87.52% accuracy rate on train data and 86.42% accuracy rate on test data. But here I can predict the ClaimStatus = 1 better with lesser error rate than all other models (~60% < ~100%)

Finally, I will check the same model with 10-Fold Cross Validation. The result is as follows.





By this method, we can predict the Specificity is higher than any other method, Accuracy too comes as approx.89%, which is not bad. Also this model minimizes class imbalance problem. So Naïve-Bayes model with 10-Fold Cross Validation is the best model.

**Recommendation:**

1. Bookmark the high valued customers in terms of high premium value, timely payment, renewal, high valued vehicle, age greater than 40years etc. to reach to them over multiple communication method to provide them the best insurance offer and coverage.
2. The policies should be made according to clientele. A complete and compact kind of plan is much welcome among the policy holders or the new customers.
3. The claim process must be hassle free and straight forward with no hidden conditions. The claiming customer must not be harassed by any means, only that way can retain the customer for lifetime.
4. Regular review of the business model is advised. This will ensure model validity or any fine tuning requirements.
5. It is also recommended to the company to increase the digital footprint to acquire new customer. Zone-wise marketing plan to be created for maximum closer.